Transfer learning of the expressivity using flow metric learning in multispeaker text-to-speech synthesis

Ajinkya Kulkarni, Vincent Colotte, Denis Jouvet

Transfer learning of the expressivity using FLOW metric learning in multispeaker text-to-speech synthesis

Ajinkya Kulkarni, Vincent Colotte, Denis Jouvet

Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France.

ajinkya.kulkarni@loria.fr, vincent.colotte@loria.fr, denis.jouvet@inria.fr

Abstract

In this paper, we present a novel flow metric learning architecture in a parametric multispeaker expressive text-to-speech (TTS) system. We proposed inverse autoregressive flow (IAF) as a way to perform the variational inference, thus providing flexible approximate posterior distribution. The proposed approach condition the text-to-speech system on speaker embeddings so that latent space represents the emotion as semantic characteristics. For representing the speaker, we extracted speaker embeddings from the x-vector based speaker recognition model trained on speech data from many speakers. To predict the vocoder features, we used the acoustic model conditioned on the textual features as well as on the speaker embedding. We transferred the expressivity by using the mean of the latent variables for each emotion to generate expressive speech in different speaker's voices for which no expressive speech data is available.

We compared the results obtained using flow-based variational inference with variational autoencoder as a baseline model. The performance measured by mean opinion score (MOS), speaker MOS, and expressive MOS shows that N-pair loss based deep metric learning along with IAF model improves the transfer of expressivity in the desired speaker's voice in synthesized speech.

Index Terms: text-to-speech, variational autoencoder, inverse autoregressive flow, deep metric learning, expressivity

1. Introduction

The parameterization of speech is still a bottleneck step in an state-of-art text-to-speech system. The development of end-to-end text-to-speech models heavily relies on encoder-decoder attention based neural network architectures which map textual vector representation to a sequence of frames of spectrograms [1–4]. Currently, the speaking style of the synthesized speech signal is neutral, as a result of the type of speech data used for training text-to-speech systems. Multispeaker expressive speech synthesis is still an open problem due to the limited availability of expressive speech corpora and time involved in the collection and annotation of such corpora for a new speaker.

Recent successes of variational inference in bayesian learning paved the way to obtain state-of-art results in various applications such as semi-supervised classification [5], generative models of images [6], voice conversion [7] and is widely used as a tool for investigating latent space for analysis of semantic information [8]. In spite of these state-of-art results, the ability of variational inference is constrained due to intractable posterior distributions to be approximated by the class of known probability distributions, over which we search for the best approximation to the true posterior [9]. The central issue in variational inference is the selection of approximate posterior distribution. In proposed inverse autoregressive flow model posterior distributions are formulated using a series of cascaded invertible transformations to map simple initial density to arbitrarily complex, flexible distribution with tractable Jocabians [10]. These cascaded transformations are called as normalizing flows.

Inverse autoregressive flow (IAF) models have been used previously in the context of speech processing applications. For fast and high-fidelity wavenet based speech synthesis, the authors in [11] proposed probability density distillation to fill in the bridge between trained Wavenet as teacher model and IAF as a student model. In [12], the authors proposed a universal audio synthesizer built using normalizing flows to learn the latent space representation for semantic control of a synthesizer by interpolation of latent variables. In this paper, IAF were used as normalizing flows to perform the variational inference for learning meaningful latent space representation .

In [13–15] the authors proposed to use reference encoder to learn disentangling in latent space, the speaker embedding which is then used to derive speech signal for the desired speaker in a tacotron based speech synthesis system. On the other hand, we opted for a few shot learning approach in which speaker information is extracted using x-vector embeddings derived from the pretrained speaker recognition model [16]. We used the extracted x-vector embeddings to train a speaker encoder network to generate speaker representation for the multispeaker TTS system.

Several approaches have been proposed previously to transfer expressivity either by controlling the prosody parameters in latent space for speech synthesis or by transferring the expressivity using interpolation of conditional embeddings of speaker identity and prosody embedding [13–15,17–19]. In [20], the authors proposed a conditional VAE (CVAE) model for expressive audio-visual synthesis. The CVAE model is constrained for single speaker audio-visual synthesis. In our work, we present x-vector based speaker embedding, which paved the way to build a multi-speaker expressive TTS system. In our approach, IAF based acoustic model is conditioned on textual features along with speaker embedding. With this conditioning, we expect to extract emotion information in latent space representation.

Recently, deep metric learning is a popular approach to train the classifiers for computer vision applications [21, 22]. Also, multiclass N-pair loss has shown superior performance compared to triplet loss or contrastive loss by considering one positive sample and $N-1$ negative samples for N classes [22]. For generating desired expressivity in synthesized speech, we need to have latent space representation clustered with respect to emotions. Thus, to disentangle latent space focusing on emotion as semantic information, we augment deep metric learning into variational inference [23]. In this paper, we presented multi class N-pair loss along with variational inference performed by IAF as a deep flow metric learning tool to enhance the latent space representation of expressivity.

This paper focuses on multiclass N-pair loss in the acous-

tic model based on IAF. We use BLSTM neural network explicitly for predicting duration for each phoneme as explained in [24]. The paper is organized as: Section 2 describes multispeaker TTS; Section 3 presents speaker embedding; Section 4 provides details about data preparation; Section 5 presents experimentation setup; Section 6 illustrates results, and Section 7 discusses conclusion.

2. Multispeaker expressive TTS

We describe our proposed IAF architecture in acoustic model and we used recurrent conditional variational autoencoder (RCVAE) model [25] as baseline system described in subsection 2.2.

2.1. Inverse autoregressive flow

The inverse autoregressive flow was introduced in 2017 [10], as a way to scale well to high-dimensional latent spaces as well as allowing faster inference. This family of flow has a series of cascaded inverse autoregressive transformation. The architecture for IAF model have three components namely encoder, IAF flow, and decoder as shown in figure 1. For given input x, encoder network generates an hidden output h. Then, hidden output h is given to feedforward neural network to obtain the μ_0 and σ_0. The initial latent variable, z_0 is estimated by drawing random sample $\varepsilon \sim \mathcal{N}(0, I)$ for using the reparameterization as shown in (1). Afterward, z_0 along with hidden output h is provided to k steps of inverse autoregressive transformation to obtain flexible posterior probability distribution with latent variable z_k, refer (2).

$$z_0 = \mu_0 + \sigma_0 \odot \varepsilon \tag{1}$$

$$z_k = \mu_k + \sigma_k \odot z_{k-1} \tag{2}$$

For each step of flow transformation, neural network based flow designed to predict μ_k and σ_k, where latent variable from the previous flow step $k-1$ and hidden output from the previous flow step are provided as an input. And amortization is performed using hidden output h as input to autoregressive networks of flow transformations [26]. These autoregressive transformations are invertible if $\sigma_i > 0$ condition is satisfied for i^{th} value of D dimension. Autoregressive structure of flow allows simple computation of the Jacobian determinant of each transformation as a change in global posterior probability density of encoder network denoted as $logQ(z_K|x)$, where z_K is output of last flow step. The equation (3) provides tractable change in probability density for which detailed derivation is provided in [10]. In this way, the flexible, tractable posterior distribution is created to perform the variational inference with the inverse autoregressive flow. Thus, ability of flexible distribution to fit closely to true posterior improves the performance of autoregressive model and depth of the chain.

$$logQ(z_K|x) = -\sum_{i=1}^{D}\left(\frac{1}{2}\varepsilon_i^2 + \frac{1}{2}log(2\pi) + \sum_{k=0}^{K} log(\sigma_{k,i})\right) \tag{3}$$

2.2. Proposed model

We implemented BLSTM layers for designing encoder and decoder of IAF acoustic model. The output of enocder network is used to estimate initial mean μ_0, initial variance σ_0, and hidden output h. Afterward, z_0 along with hidden output h is given to IAF transformation to obtain z_k after k transformations. We have conditioned the decoder network with flow transformed z_k along with condition c corresponding to textual features, duration information, and speaker embedding. The decoder network generates predicted acoustic features $\hat{x}$ as a output. During the training phase, $\hat{x}$ is then used for computing the reconstruction loss, $logP(x|z,c)$. In the inference phase, we sample z_0 from latent space to obtain z_k after flow transformation. Then z_k is given to the decoder network with condition c, to obtain acoustic features for speech synthesis using a vocoder. Therefore, choosing an appropriate latent variable is a crucial factor in generating appropriate expressivity in synthesized speech.

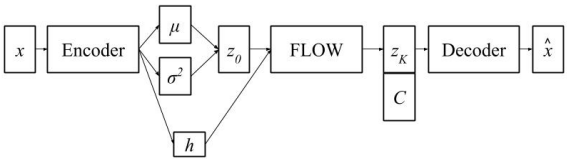

Figure 1: *Inverse autoregressive flow model for acoustic model in parameteric expressive multispeaker TTS.*

$$\text{Loss} = \mathrm{E}_z[logP(x|z,c)] + \lambda \cdot logQ(z_K|x) + \beta \cdot log(1 + \sum_{i=1}^{N-1} exp(z^\top z_i^- - z^\top z^+)) \tag{4}$$

We propose to add a multi-class N-pair loss criteria as deep metric learning to variational inference along with other losses. The multi-class N-pair loss enhances the latent representation compared to triplet loss by pushing away multiple negative samples at each backpropagation update step [22]. This results in increasing the intercluster distance from $N-1$ negative samples and decreases the intracluster distance between positive samples and training examples. In this case, N refers to number of emotions, positive samples refer to latent variables from the same emotion class and negative samples correspond to latent variables sampled from different emotion classes.

The proposed multi-class N-pair loss function is applied to the initial latent variable z_0 before the flow transformation step. For N classes, z^+ is a positive sample and $\{z_i^-\}_1^{N-1}$samples are from negative classes as stated in equation (4). For sampling positive sample and negative samples, we used the precomputed mean of latent variables for each emotion. The figure 2 b. and figure 2 d., clearly illustrates the improvement of latent representation of emotions after addition of multiclass N-pair loss. During the inference phase, means of latent variables for each emotions are pre-computed. This pre-computed mean is provided to decoder network to transfer the expressivity in desired speaker's voice.

2.3. Baseline model

In this paper, we are investigating IAF model and comparing baseline model based on recurrent conditional variational autoencoder (RCVAE) [25] with IAF acoustic model. We modified the conventional RCVAE architecture to handle a speaker embedding for multispeaker TTS as well as multiclass n-pair loss as a deep metric learning tool for enhancing the latent representation of expressivity as a semantic information.

Variational autoencoders have components such as encoder, decoder and loss function. The loss function corresponds to

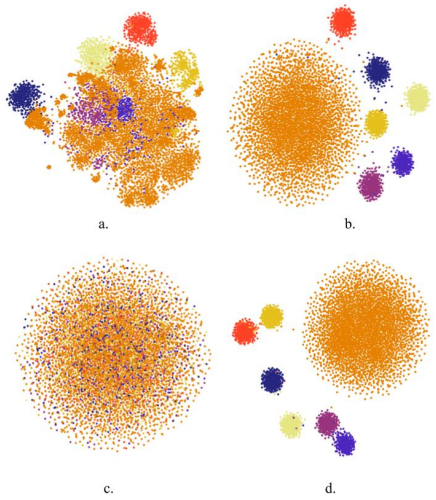

Figure 2: *t-SNE plot of latent representation of RCVAE acoustic model (a.), and RCVAE acoustic model with N-pair loss (b.); IAF acoustic model (c.), IAF acoustic model with N-pair loss (d.);Each color in t-SNE plot represents emotion; Here, neutral style by several speakers is represented by orange*

the reconstruction loss plus a regularization term defined with a Kullback-Leibler (KL) divergence. Furthermore, we augmented multiclass n-pair loss for clustering of latent variables of each emotion. We provided KL divergence with RCVAE framework as a means to evaluate of variational inference done by flow based architecture. We substituted the KL divergence loss term with change in probability density, $logQ(z_K|x)$. Besides this, all the implementation details for RCVAE model are exactly same including addition of multiclass N-pair loss and transferring expressivity using pre-computed latent variables, as explained in subsection 2.2.

3. Speaker Embedding

We proposed to develop a speaker encoder network using x-vector embeddings extracted from a pretrained speaker recognition model. The x-vector embeddings are deep neural network based embeddings trained on time-delay neural networks with a statistical pooling layer trained for the speaker recognition task [16]. Firstly, we extracted x-vector embeddings from the pretrained speaker recognition model trained on the voxceleb corpus available in the Kaldi toolkit [27, 28].

To adapt the speaker embeddings to French speakers, we used extracted x-vector embeddings to train a feedforward neural network based speaker recognition model for discriminating between the speakers of our French speech synthesis corpora. Even though speaker encoder is not trained to capture speaker identity directly, experimentation with speaker embeddings shown capablity to represent speaker characteristics in synthesized speech.

4. Data preparation

We used 4 speech corpora for developing our multispeaker expressive TTS system. The speech corpora are Lisa [12], a French female neutral corpus (approx. 3 hrs), Caroline [20], a French female expressive corpus (approx. 9hrs), SIWIS [29], a French female neutral corpus (approx. 3 hrs), and Tundra [30], a French male neutral corpus (approx. 2hrs). Beside neutral speech, Caroline's expressive speech corpus has 6 emotions namely joy, surprise, fear, anger, sadness, and disgust (approx. 1hr for each emotion and 3hrs for neutral). All the speech signals were used at a sampling rate of 16 kHz. Each speech corpus is divided into train, validation and test sets in the ratio of 80%, 10%, 10% respectively.

We parameterized speech using the WORLD vocoder [31] with 187 acoustic features computed every 5 milliseconds, namely 180 spectral features as Mel generalized cepstrum coefficients (mgc), 3 log fundamental frequencies (lf0), 3 band-aperiodicities (bap) and 1 value for voiced-unvoiced information (vuv). We applied z-normalization on extracted acoustic features from the WORLD vocoder. For converting French text to linguistic features (also known as context labels, dimension 180), SOJA-TTS tool (internally developed in our team) is used as a front-end text processor.

5. Experimentation

For implementing IAF and RCVAE architectures, we built the encoder network and the decoder network using 2 BLSTM layer of 256 hidden units each, a latent variable of 50 dimensions, a learning rate of 0.001, Adam optimizer initialized with default parameters and a batch size of 10. For training the RCVAE model, we added the λ of 0.001 to the KL divergence loss term, and β multiplication factor of 1.0 is added to the multiclass N-pair loss term. For IAF model, we used β of value 1, while λ factor of 0.025 is added to the multiclass N-pair loss term and incremented with 0.025 with each epoch. Both the baseline model and the IAF model are trained for 50 epochs. To ensure better convergence of model parameters multi-class N-pair loss was activated only after the first 5 epochs. In the inference phase, we used the mean of latent variables constituting given emotions as a latent variable to synthesize a particular emotion.

We implemented 5 layers of multilayer perceptron trained to discriminate 4 French speakers (corresponding to our speech synthesis corpora) with 512-256-128-64-16 hidden units. We provided input as extracted 512-dimensional x-vector to multilayer perceptron network. The network is trained using cross-entropy loss criteria, Adam optimizer, and 50 epochs of training. The speaker embeddings are generated by extracting the output of the last hidden layer of dimension 16.

6. Results

We computed mel cepstrum distortion (MCD) for objective evaluation of presented models for all the speakers as shown in Table 1. Furthermore, we conducted the Mean Opinion Score (MOS) [32] perception test for subjective evaluation of proposed IAF based multi-speaker expressive text-to-speech synthesis system as well as for the RCVAE model. In the MOS perception test, each listener rated synthesized speech stimuli on a scale from 1 to 5 score, where 1 is bad and 5 is excellent, considering intelligibility, naturalness, and quality of the speech stimuli. The 12 French listeners participated in the perception test; each listener had to score 5 stimuli for each speaker-emotion pair randomly chosen from the test set. The results of the MOS test are shown in Table 2 with an associated 95% confidence intervals.

The listening experiment with RCVAE model shown that addition of N-pair loss significantly improves the speaker MOS and expressive MOS. An informal listening experiment leads to similar conclusion for IAF model and also figure 2 c. shows no visible clusters of emotion. Thus, we opted out IAF model

without N-pair loss from the subjective evaluation. The IAF N-pair model shown better MOS values than other models, except Lisa speaker's voice for which RCVAE N-pair model scored 3.1. The MOS score presented for Caroline speaker in Table 2 represents the average score obtained for Caroline's neutral voice and for all Caroline emotions. Due to limited training data (1hr) for each emotion for Caroline's voice, MOS score performance on Caroline's speech synthesis is lower compared to other speakers.

We proposed speaker MOS and expressive MOS to evaluate the performance of the presented architecture for transfer of expressivity onto other speaker voices. In speaker MOS test, we directed listeners to assign a score regards to the similarity between a reference speaker speech stimuli and synthesized expressive speech in a range of 1 (bad) to 5 (excellent). Similarly, we also instructed listeners to score expressivity perceived in the synthesized expressive speech on a scale of 1 (bad) to 5 (excellent). The score is assigned depending on the closeness of expressive characteristics in synthesized speech compared to reference expressive speech stimuli. 12 French listeners participated in a perception test, each listener scored 3 sets of stimuli for each target speaker-emotion pair. The results of expressive MOS and speaker MOS are shown in Table 3 and Table 4, with associated 95% confidence intervals.

Table 1: *Objective evaluation using MCD results*

Model	MCD
RCVAE	5.795
RCVAE+N-pair	5.472
IAF+N-pair	**5.144**

Table 2: *MOS score for evaluation of Multi-speaker TTS system*

MOS	Caroline	Lisa	Siwis	Tundra
RCVAE	2.4 ± 0.3	2.8 ± 0.7	2.6 ± 0.8	2.7 ± 0.2
RCVAE+N-pair	2.9 ± 0.2	**3.1 ± 0.6**	3.0 ± 0.5	2.9 ± 0.4
IAF+N-pair	**2.9 ± 0.3**	3.0 ± 0.5	**3.2 ± 0.4**	**3.0 ± 0.6**

Table 3: *Speaker MOS score for evaluation of transfer of speaker characteristics*

Speaker MOS	Lisa	Siwis	Tundra
RCVAE	2.3 ± 0.2	2.2 ± 0.1	2.7 ± 0.3
RCVAE+N-pair	3.0 ± 0.1	2.7 ± 0.3	2.9 ± 0.2
IAF+N-pair	**3.0 ± 0.3**	**2.8 ± 0.2**	**3.0 ± 0.4**

From figure 2, t-SNE representation of IAF N-pair model have tighly bounded clusters compared to RCVAE N-pair model which have more outliers for cluster of emotions. This is inline with results obtained for speaker MOS, and expressive MOS, where IAF N-pair model slightly performed better than RCVAE N-pair model. The RCVAE model without N-pair loss performed poorly compared to other models. The speaker MOS, as well as Expressive MOS, showed that while transferring expressivity, the addition of N-pair loss improves the retainment of the speaker's characteristics. Furthermore, the presented approach was equally able to transfer the expressivity from female (Caroline) to female (Lisa, Siwis) speakers as well as female (Caroline) to male (Tundra) speaker. This shows that variational inference performed using IAF models improves the perceived expressivity in the desired speaker's voice as a result of the flexible, tractable posterior distribution.

Table 4: *Expressive MOS score for evaluation of transfer of expressivity*

Expressive MOS	Lisa	Siwis	Tundra
RCVAE	1.4 ± 0.4	1.5 ± 0.3	1.7 ± 0.5
RCVAE+N-pair	1.9 ± 0.3	1.9 ± 0.4	2.0 ± 0.2
IAF+N-pair	**2.1 ± 0.2**	**2.0 ± 0.3**	**2.0 ± 0.4**

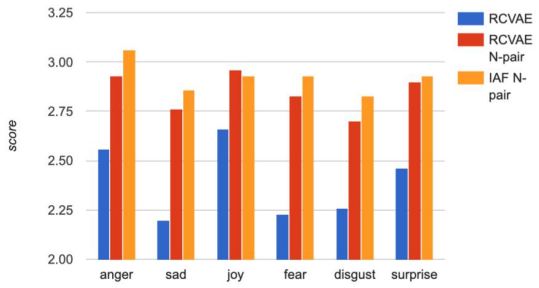

Figure 3: *Speaker MOS score*

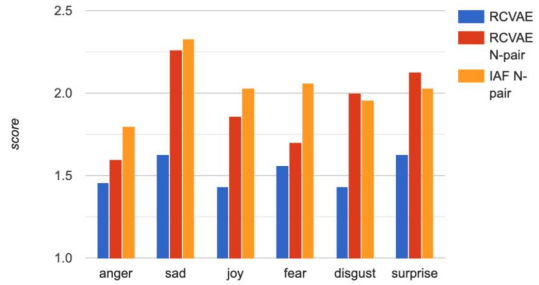

Figure 4: *Expressive MOS score*

7. Conclusion

We presented inverse autoregressive flow model with deep metric learning to transfer the expressivity to desired speaker's voice in multispeaker text-to-speech synthesis system. In our approach, the deep flow metric learning helped to enforce the better clustering of emotions in latent space representation. The presented work is the first approach that uses deep metric learning in an inverse autoregressive flow based variational inference. The obtained results show that the proposed framework enhances latent space representation in a multi-speaker expressive TTS system.

Thereafter, from the MOS scores, the x-vector based speaker embedding helps multispeaker TTS system to represent the speaker characteristics. The subjective evaluation conducted show that the proposed approach enhances ability to transfer the expressivity. In the future, we would like to implement a similar latent space representation in end-to-end TTS system for adapting expressivity in new speaker's voice.

8. Acknowledgements

Experiments presented in this paper were carried out using the Grid'5000 testbed, supported by a scientific interest group hosted by Inria and including CNRS, RENATER and several Universities as well as other organizations (see https://www.grid5000.fr).

9. References

[1] Y. Wang, R. J. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. V. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: A fully end-to-end text-to-speech synthesis model," *ArXiv*, vol. abs/1703.10135, 2017.

[2] W. Ping, K. Peng, A. Gibiansky, S. Ö. Arik, A. Kannan, S. Narang, J. Raiman, and J. L. Miller, "Deep voice 3: 2000-speaker neural text-to-speech," *ArXiv*, vol. abs/1710.07654, 2017.

[3] J. Sotelo, S. Mehri, K. Kumar, J. F. Santos, K. Kastner, A. C. Courville, and Y. Bengio, "Char2wav: End-to-end speech synthesis," in *ICLR*, 2017.

[4] Y. Taigman, L. Wolf, A. Polyak, and E. Nachmani, "Voiceloop: Voice fitting and synthesis via a phonological loop," in *ICLR*, 2017.

[5] D. P. Kingma, S. Mohamed, D. J. Rezende, and M. Welling, "Semi-supervised learning with deep generative models," in *NIPS*, 2014.

[6] M. El-Kaddoury, A. Mahmoudi, and M. M. Himmi, "Deep generative models for image generation: A practical comparison between variational autoencoders and generative adversarial networks," in *Mobile, Secure, and Programmable Networking*, 2019, pp. 1–8.

[7] A. van den Oord, O. Vinyals, and K. Kavukcuoglu, "Neural discrete representation learning," in *NIPS*, 2017.

[8] Y. Bengio, A. C. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1798–1828, 2013.

[9] D. J. Rezende and S. Mohamed, "Variational inference with normalizing flows," *ICML*, pp. 1730–1538, 2015.

[10] D. P. Kingma, T. Salimans, R. Jozefowicz, X. Chen, I. Sutskever, and M. Welling, "Improved variational inference with inverse autoregressive flow," in *Advances in Neural Information Processing Systems*, 2016, pp. 4743–4751.

[11] A. van den Oord, Y. Li, I. Babuschkin, K. Simonyan, O. Vinyals, K. Kavukcuoglu, G. van den Driessche, E. Lockhart, L. C. Cobo, F. Stimberg, N. Casagrande, D. Grewe, S. Noury, S. Dieleman, E. Elsen, N. Kalchbrenner, H. Zen, A. Graves, H. King, T. Walters, D. Belov, and D. Hassabis, "Parallel wavenet: Fast high-fidelity speech synthesis," *ICML*, pp. 3915–3923, 2018.

[12] P. Esling, N. Masuda, A. Bardet, R. Despres, and A. Chemla-Romeu-Santos, "Universal audio synthesizer control with normalizing flows," *ArXiv*, vol. abs/1907.00971, 2019.

[13] W.-N. Hsu, Y. L. Zhang, R. J. Weiss, H. Zen, Y. Wu, Y. Wang, Y. Cao, Y. Jia, Z. Chen, J. Shen, P. Nguyen, and R. Pang, "Hierarchical generative modeling for controllable speech synthesis," 2019.

[14] Y. Wang, D. Stanton, Y. Zhang, R. J. Skerry-Ryan, E. Battenberg, J. Shor, Y. Xiao, F. Ren, Y. Jia, and R. A. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," 2018.

[15] R. J. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, R. J. Weiss, R. Clark, and R. A. Saurous, "Towards end-to-end prosody transfer for expressive speech synthesis with tacotron," *ArXiv*, pp. 4700–4709, 2018.

[16] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," *ICASSP*, pp. 5329–5333, 2018.

[17] Y.-J. Zhang, S. Pan, L. He, and Z.-H. Ling, "Learning latent representations for style control and transfer in end-to-end speech synthesis," *ICASSP*, pp. 6945–6949, 2019.

[18] K. Akuzawa, Y. Iwasawa, and Y. Matsuo, "Expressive speech synthesis via modeling expressions with variational autoencoder," in *Interspeech*. ISCA, 2018, pp. 3067–3071.

[19] Y. Lee and T. Kim, "Robust and fine-grained prosody control of end-to-end speech synthesis," *ICASSP*, pp. 5911–5915, 2019.

[20] S. Dahmani, V. Colotte, V. Girard, and S. Ouni, "Conditional variational auto-encoder for text-driven expressive audiovisual speech synthesis," in *INTERSPEECH*, 2019.

[21] M. Kaya and H. Şakir Bilge, "Deep metric learning: A survey," *Symmetry*, vol. 11, p. 1066, 2019.

[22] K. Sohn, "Improved deep metric learning with multi-class n-pair loss objective," in *NIPS*, 2016.

[23] X. Lin, Y. Duan, Q. Dong, J. Lu, and J. Zhou, "Deep variational metric learning," in *ECCV*, 2018.

[24] Z. Wu, O. Watts, and S. King, "Merlin: An open source neural network speech synthesis system," in *SSW*, 2016.

[25] A. Kulkarni, V. Colotte, and D. Jouvet, "Deep Variational Metric Learning For Transfer Of Expressivity In Multispeaker Text To Speech," in *SLSP*, 2020.

[26] R. van den Berg, L. Hasenclever, J. M. Tomczak, and M. Welling, "Sylvester normalizing flows for variational inference," in *UAI*, 2018.

[27] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *INTERSPEECH*, 2018.

[28] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. K. Goel, M. Hannemann, P. Motlícek, Y. Qian, P. Schwarz, J. Silovský, G. Stemmer, and K. Veselý, "The kaldi speech recognition toolkit," 2011.

[29] J. Yamagishi, P.-E. Honnet, P. N. Garner, and A. Lazaridis, "The siwis french speech synthesis database," 2017.

[30] A. Stan, O. Watts, Y. Mamiya, M. Giurgiu, R. A. J. Clark, J. Yamagishi, and S. King, "Tundra: a multilingual corpus of found data for tts research created with light supervision," in *INTERSPEECH*, 2013.

[31] M. Morise, F. Yokomori, and K. Ozawa, "World: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE*, pp. 1877–1884, 2016.

[32] R. C. Streijl, S. Winkler, and D. S. Hands, "Mean opinion score (mos) revisited: methods and applications, limitations and alternatives," *Multimedia Systems*, pp. 213–227, 2014.